Unified Multimodal Models as Auto-Encoders

Supplementary Material

Supplementary Overview

- Section 1: Additional Related works.
- Section 2: Dataset details.
- Section 3: Training settings.
- Section 4: Qualitative examples.
- Section 5: Additional experimental results.

1. Additional Related works

Reinforcement Learning in Generative Models. The widespread success of Reinforcement Learning from Human Feedback (RLHF) in aligning large language models (LLMs) with human intent [9, 14] has inspired its application to text-to-image generation. In this context, a common strategy involves first training a reward model (RM) that learns from human judgments—either general aesthetic preferences [46] or alignment between prompts and generated images [45], followed by reinforcement learning to optimize the generative model accordingly [4]. Despite its promise, this two-stage approach faces significant limitations when applied to image editing tasks. Reward models are often brittle and challenging to design robustly [29], and they can be gamed through superficial changes that maximize reward without improving actual quality, a phenomenon known as ”reward hacking” [41]. More recently, alternative optimization frameworks like GRPO [38] have emerged as viable solutions, demonstrating effectiveness in tuning both diffusion and flow-matching based models. Extensions such as FlowGRPO [26] and DanceGRPO [47] illustrate the adaptability of these algorithms to complex generative processes, offering a more stable and fine-grained path toward aligning visual outputs with human expectations—particularly in dynamic, iterative editing scenarios where traditional methods fall short.

Benchmarking Multimodal Understanding, Generation, and Unification. Evaluating unified multimodal models (UMMs) typically involves aggregating performance across multiple specialized benchmarks, each targeting distinct capabilities. For assessing visual understanding, widely adopted benchmarks include ScienceQA [28], MMMU [49], VQA [2], GQA [17], and MM-Bench [27], all of which rely heavily on large-scale datasets with human-annotated images and labels. In contrast, our proposed UniBench introduces a novel paradigm as a VQA-style benchmark specifically designed for *generated* images, eliminating the dependency on real-image annotations by evaluating comprehension directly on synthesized content. For generative capability assessment, image quality is commonly measured using metrics such as FID [13], ImageReward [45], and LIQ [39], often evaluated on standard image corpora like MSCOCO [25] or LAION-5B [37]. Additional factors such as text-image alignment [12], fairness [19], and stylistic consistency [32] are also considered, drawing from benchmarks like HRS [3]. However, unified models place greater emphasis on instruction-following and coherent joint reasoning across perception and generation. As such, evaluation frameworks tailored to text-to-image synthesis, such as GenEval [11], DPG-Bench [15], and T2I-CompBench++ [16], which are particularly relevant. These assess fine-grained attributes including object presence, spatial relations, counting accuracy, color fidelity, and positional reasoning [3, 8, 20]. Despite their utility, existing benchmarks are not specifically designed for the dual perception-generation nature of UMMs, leaving a gap in comprehensive, integrated evaluation. To address world-knowledge grounding in image synthesis, WISE [30] was recently introduced to evaluate models' implicit understanding of real-world constraints across domains such as food preparation, material physics, and object affordances. More recently, UniEval [22] proposes a new benchmark dedicated to unified multimodal modeling, covering a broader range of semantic, structural, and logical challenges with increased task difficulty and potential for model improvement.

2. Dataset Details

RL stage data (1K). For the reinforcement learning (RL) phase, we curate a compact but highly refined dataset of 1,000 real-world photography images selected for exceptional compositional quality, visual clarity, and semantic richness. These images span diverse domains such as portrait photography, architectural shots, nature scenes, and dynamic street photography, all captured under realistic lighting and perspective conditions. In addition to these hand-picked photographs, we incorporate a specialized subset of synthetic yet photorealistic data from Echo-4o [48], which provides tightly aligned text-image pairs with expert-level captions and controlled visual variations. This combined RL dataset is used in a reconstruction-driven optimization framework: given a caption derived from one of these target images, the model is tasked with generating a new image, and its output is evaluated against the original using a learned reward model that assesses fidelity, detail preservation, and semantic alignment. Through this closed-loop paradigm, improved captioning leads to better reconstruction, which in turn refines generation capabilities.

Data for evaluation in Unified-Bench. To evaluate the model's performance on the proposed Unified-Bench, we randomly sample 100 images from the LAION-5B dataset [37] to serve as a dedicated test split. These images are selected without any filtering or curation based on content or aesthetic score, ensuring a representative and unbiased distribution across categories, styles, and complexity levels.

3. Training Settings

Training details of Unified-GRPO. We employ the GRPO RL algorithm [38] to fine-tune only the LLM module while keeping other modules, like the corresponding visual encoder/decoder, frozen. We empirically observe that updating the visual encoder during RL training can lead to instability and degradation in image quality (see Fig. 1), such as anomaly artifacts, structural collapse, or semantic inconsistency, so we disable its gradient updates to preserve visual feature integrity. To enable effective sampling for RL-based image generation, we treat the combination of the diffusion decoder and a pre-trained CLIP model [35] as a unified, frozen reward module. This composite model operates purely in inference mode: given a generated image and its corresponding reconstructed image from the input caption, it computes a similarity score that serves as the final reward signal in the GRPO framework. During training, we use a learning rate of 1×10^{-6} and a batch size of 1 due to the high computational cost of diffusion-based RL. For each prompt, we generate 4 sampled images to estimate the policy gradient in GRPO, and we set the KL regularization coefficient 1×10^{-6}, indicating that we only apply a minimal penalty for divergence from the reference policy, focusing solely on reward maximization. The temperature of LLM is set to be 1.0. The prompt used to do the LLM sampling is shown below (see Prompt. 5).

Note that we do **not** explicitly require the LLM to generate descriptive or comprehensive captions during training. After RL, the LLM autonomously produces longer and richer captions that are more conducive to high-fidelity image generation, even though no explicit supervision or loss is applied to the caption content itself. This emergent behavior suggests that the ***RL signal from image reconstruction quality implicitly guides the LLM toward generating more detailed and image-friendly textual descriptions.***

4. Qualitative Examples

Enhancing model's comprehensive perception by the generation model. Fig. 3 contrasts captions used for reconstruction on a challenging example (small black dog wearing a yellow beanie and glasses). Baselines reveal three typical errors. (i) *Category drift*: some misidentify the subject as a monkey, causing the generator to synthesize an incorrect species. (ii) *Attribute omissions or swaps*: descriptions drop key items (beanie, glasses) or mismatch apparel colors, leading to reconstructions that caricature the outfit. (iii) *Under-specified scenes*: vague backgrounds and missing lighting cues prevent consistent photographic style at inference. UAE's caption, in contrast, enumerates the full set of semantics—species, apparel *type and color*, eyewear, pose, occlusions ("ears are not visible"), background style ("blurred, park-like"), and lighting—producing a reconstruction that preserves identity, attire, and overall aesthetic. This example typifies the mechanism by which better understanding (denser, better-bound descriptions) yields better generation, echoing our Unified-Bench gains in Tab. **??**.

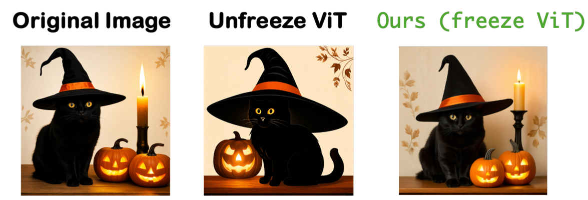

Figure 1. Illustration of the reconstruction results when unfreezing ViT (the visual encoder of the MLLM) for joint training. We observe that the generated output collapses, semantically important details such as "two pumpkins" and "one candle" are missing. This degradation motivates us to keep the ViT frozen during finetuning across all experiments.

5. Additional Experimental Results

The text-to-image generation results on DPG-Bench. On DPG-Bench (Tab. 1), UAE achieves the top scores on *Entity* (91.43), *Attribute* (91.49), and *Relation* (92.07), and ranks second overall with **84.74**, closely trailing Bagel (85.07). The sub-score pattern suggests UAE's advantages come from faithful entity grounding and relation handling under long prompts, translating into competitive end-to-end generation quality within a unified architecture.

Prompt list used in Fig. 2. We provide the full caption for each sample in generation order, reading from left to right and top to bottom, row by row.

- **Sample-1**. A close-up portrait of a ginger tabby cat, its fur a rich tapestry of warm amber and deep russet stripes that catch the soft, directional light illuminating its face from the side, highlighting the velvety texture of its coat and the subtle contours of its cheekbones, while its large, luminous green eyes gaze intently off-camera with an expression of quiet contemplation and alert curiosity, framed by long, delicate white whiskers and perked ears that suggest attentiveness, all set against a dark, shadowy background that isolates the feline subject and enhances the dramatic, almost painterly quality of the image, emphasizing the cat's regal poise and enigmatic presence.

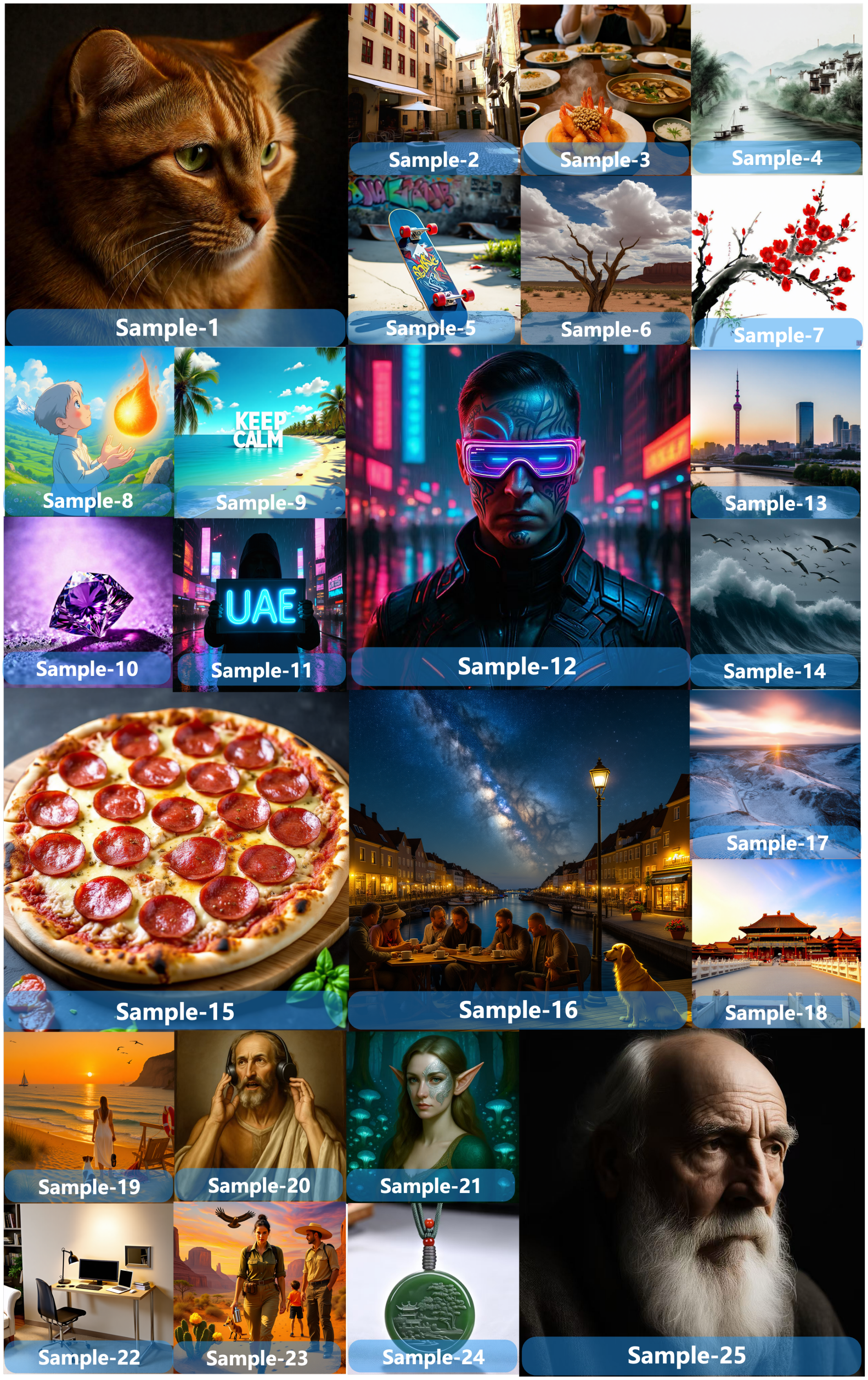

Figure 2. Visualization results of `UAE` at 1024×1024 resolution.

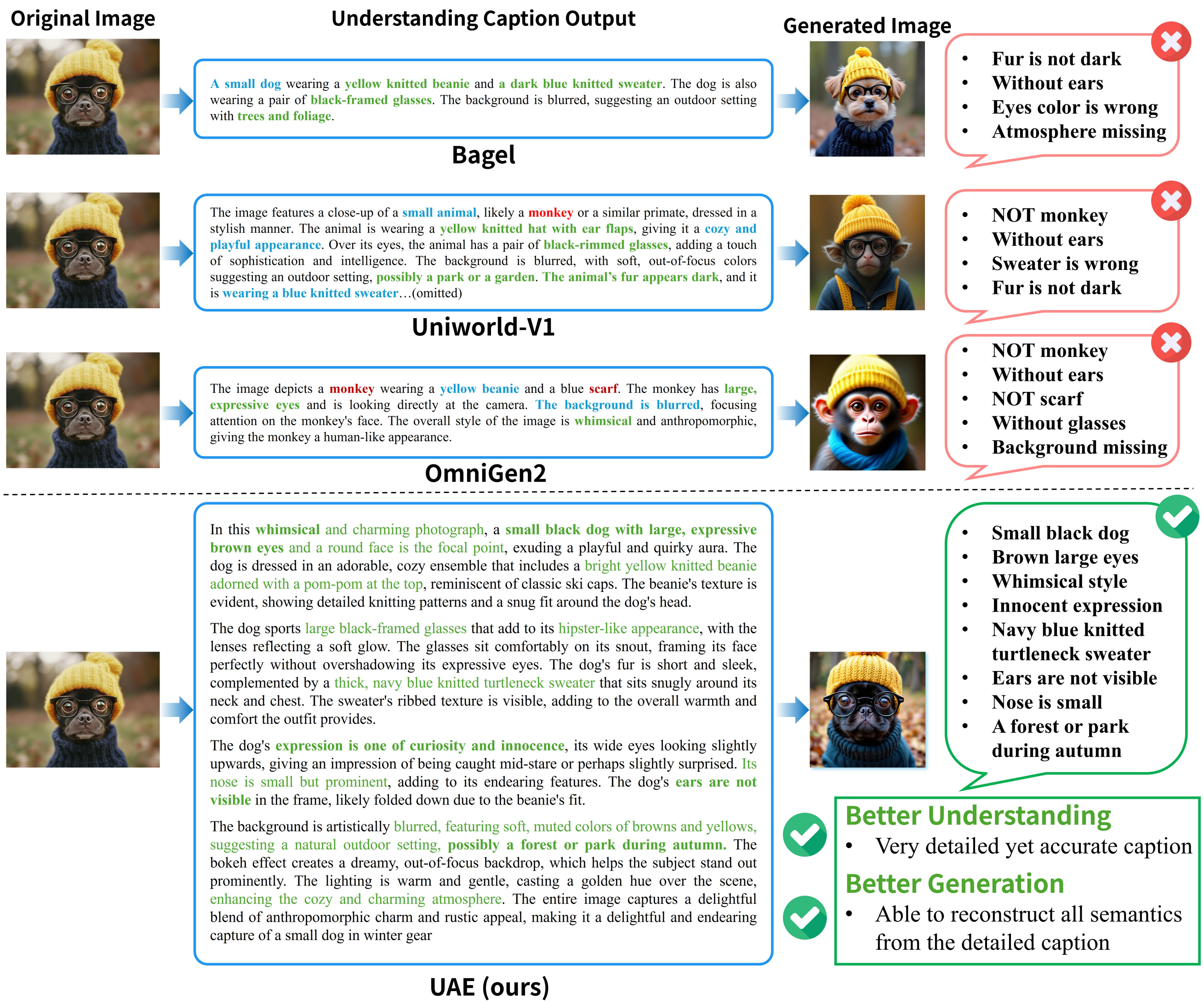

Figure 3. **Case study of the results from the proposed Unified-Bench**, we see that *our UAE enables to produce a more detailed, accurate, comprehensive description based on the input image*, and reconstructs a similar result to the original image, showcasing the improved understanding and generation capabilities, and the better unification of the system.

- **Sample-2**. The building on the left is a light beige color with a series of rectangular windows framed in red, some with small white panes. These windows have simple brick or mortar surrounds and are uniformly spaced, creating a rhythmic pattern across the facade. The ground floor features a small shop area with a white canopy providing shade for outdoor seating. The canopy is supported by metal poles and holds a few tables under its shelter. Behind this canopy, various items can be seen, including a few chairs and tables, indicating a café or small eatery. A white umbrella stands next to the shop entrance, adding to the cozy atmosphere.Above the shop, the building has a series of small balconies with metal railings, each adorned with potted plants and hanging baskets, contributing to the pedestrian-friendly urban design. The ground floor has a mix of business signs, some of which are partially visible but not legible, suggesting a bustling commercial area. There's a dark green signboard affixed to one of the windows, possibly indicating a specialty shop or restaurant. The neighboring building on the right is a lighter shade of beige with a pastel green section near the top. Its windows are similarly framed in red, with larger panes and a more varied arrangement compared to the first building. This building features balconies with metal railings and small rectangular windows. The exterior walls show some wear and tear, with subtle moldings and patches of weathering, adding character to the structures. In front of these buildings lies a cobblestone street, which is partially shaded by

Table 1. Comparisons of text-to-image generation ability on DPG-Bench [15] benchmark. **Bold** indicates the best result, and <u>underlined</u> denotes the second best.

Method	Global	Entity	Attribute	Relation	Other	Overall
Dedicated T2I						
SDXL [33]	83.27	82.43	80.91	86.76	80.41	74.65
PlayGroundv2.5 [21]	83.06	82.59	81.20	84.08	83.50	75.47
Hunyuan-DiT [23]	84.59	80.59	88.01	74.36	86.41	78.87
PixArt-Σ [5]	86.89	82.89	88.94	86.59	87.68	80.54
DALLE3 [31]	**90.97**	89.61	88.39	90.58	89.83	83.50
SD3-medium [1]	87.90	<u>91.01</u>	88.83	80.70	88.68	84.08
FLUX.1-dev [18]	82.1	89.5	88.7	<u>91.1</u>	89.4	84.0
OmniGen [43]	87.90	88.97	88.47	87.95	83.56	81.16
Unified Model						
Show-o [44]	79.33	75.44	78.02	84.45	60.80	67.27
EMU3 [40]	85.21	86.68	86.84	90.22	83.15	80.60
TokenFlow-XL [34]	78.72	79.22	81.29	85.22	71.20	73.38
Janus Pro [7]	86.90	88.90	89.40	89.32	<u>89.48</u>	84.19
BLIP3-o 4B [6]	-	-	-	-	-	79.36
BLIP3-o 8B [6]	-	-	-	-	-	81.60
UniWorld-V1 [24]	83.64	88.39	88.44	89.27	87.22	81.38
OmniGen2 [42]	88.81	88.83	90.18	89.37	**90.27**	83.57
BAGEL [10]	<u>88.94</u>	90.37	<u>91.29</u>	90.82	88.67	**85.07**
UAE	83.11	**91.43**	**91.49**	**92.07**	84.32	<u>84.74</u>

the shadows cast by the buildings. A large stone fountain occupies the foreground, its base circular and gray, with a worn, dark surface. The pavement around the fountain is paved with irregularly shaped stones, creating a rustic, old-world feel. The sunlight creates dramatic contrasts, with deep shadows and bright highlights accentuating the textures of the buildings and the cobblestones. The street is quiet, devoid of people, which enhances the serene and timeless atmosphere of the scene.

- **Sample-3**. A photo of hearty Chinese meal.
- **Sample-4.** This serene watercolor painting evokes the tranquil spirit of a traditional Chinese riverside village, where mist-laden mountains recede into a soft, pale sky, their layered silhouettes rendered in gentle washes of gray and muted green that dissolve into atmospheric haze; along the calm, reflective riverbank, white-walled houses with dark-tiled, upturned eaves nestle among lush trees, their architecture echoing classical Jiangnan aesthetics, while two slender wooden boats glide silently on the glassy water—one closer to the foreground with its simple mast and open cabin, the other a distant speck fading into the fog—imbuing the scene with quiet movement and timeless stillness, as the interplay of light and shadow across the rippling surface and the subtle gradations of ink suggest not only depth and distance but also a meditative harmony between nature and human habitation, capturing the essence of poetic rural life suspended in a dreamlike, almost ethereal moment.
- **Sample-5**. A vibrant blue skateboard with bold, graffiti-style graphics—featuring swirling red and yellow patterns and stylized lettering—stands upright on cracked concrete, its bright red wheels and silver trucks catching the sunlight, casting a sharp shadow on the ground, while in the blurred background, a weathered wall adorned with colorful street art and a partially visible skate ramp hint at an urban skate park setting, blending raw energy with artistic expression under a clear, sunlit sky.
- **Sample-6**. A solitary, gnarled tree with twisted, leafless branches stretches skyward like a skeletal sentinel in the heart of a vast desert landscape, its weathered trunk rooted

firmly in the ochre sands that stretch to the horizon, dotted sparsely with low-lying shrubs; above, a dramatic expanse of billowing cumulus clouds drifts across a brilliant blue sky, casting shifting shadows over the arid terrain, while in the distance, the imposing silhouette of red rock mesas rises majestically against the horizon, lending a sense of ancient grandeur and timeless solitude to the scene, where nature's raw resilience and stark beauty are captured in perfect harmony under the vast, open heavens.

- **Sample-7**. A striking traditional East Asian ink painting captures the vibrant essence of a blossoming plum tree, its gnarled, darkly rendered branches—executed with bold, expressive brushstrokes of sumi ink—arching gracefully across the stark white paper to cradle clusters of vivid crimson flowers, each petal delicately shaped with fluid washes of red that convey both vitality and fragility, while subtle hints of green foliage at the lower left suggest the quiet emergence of new life; the composition balances dynamic movement with serene stillness, evoking themes of resilience and renewal as the blossoms defiantly bloom against the void, enhanced by the faint calligraphic inscription near the trunk and the small red seal in the corner, which together anchor the piece in cultural tradition and artistic intention.
- **Sample-8**. In a breathtaking, sun-drenched meadow of lush rolling hills dotted with wildflowers and scattered boulders, a young boy with soft silver-gray hair and wide, awestruck blue eyes gazes upward in wonder as he gently cradles a radiant, living flame between his outstretched palms—a glowing, teardrop-shaped orb of golden-orange fire that pulses with warmth and light, its edges flickering with delicate embers against the backdrop of a brilliant blue sky streaked with fluffy white clouds and distant snow-capped mountains; dressed in a simple light-blue jacket over a crisp white shirt, the child embodies innocence and quiet awe, as if he has just summoned or discovered this mystical force, transforming the idyllic pastoral landscape into a realm where magic feels not only possible but tenderly held, evoking a sense of harmony between nature, wonder, and the boundless imagination of youth.
- **Sample-9**. A vibrant, sun-drenched tropical beach unfolds under a brilliant azure sky dotted with fluffy white clouds, where the crystal-clear turquoise waters gently lap against golden sands lined with swaying palm trees casting dappled shadows on the shore, and at the heart of this serene paradise, the bold, three-dimensional white letters spelling "KEEP CALM" rise majestically from the sea's edge, their clean, modern font contrasting with the organic beauty of nature while reinforcing the tranquil mood, as if the very landscape itself is whispering a soothing mantra of peace, relaxation, and escape from the chaos of everyday life.
- **Sample-10**. A dazzling, multifaceted purple diamond rests regally upon a shimmering bed of iridescent violet sand, its precisely cut facets catching and refracting beams of ethereal light that radiate from behind, casting a luminous glow across the scene and accentuating the gem's deep amethyst hue with flashes of electric violet and cool silver highlights; the background dissolves into a dreamy, softly diffused gradient of lavender and indigo, enhancing the jewel's otherworldly brilliance and making it appear almost suspended in a mystical twilight realm, where every angle of its polished surface seems to whisper secrets of rare beauty and enchanted allure, evoking both luxury and fantasy in a single, captivating moment.
- **Sample-11**. In a rain-slicked, neon-drenched cyberpunk cityscape at night, a mysterious hooded figure stands silhouetted against a kaleidoscope of glowing skyscrapers and pulsating billboards, their face obscured by shadow as they hold aloft a luminous rectangular sign that boldly proclaims "UAE" in vibrant, electric-blue neon lettering, casting an otherworldly glow on their gloved hands and the wet pavement below, where reflections of magenta, cyan, and violet lights ripple across the glossy street like liquid electricity, evoking a futuristic vision of the United Arab Emirates as a nexus of technology, mystery, and urban energy under a dark, rain-streaked sky.
- **Sample-12**. A cybernetic warrior stands resolute in the heart of a rain-lashed, neon-soaked metropolis, his face etched with intricate biomechanical tattoos that glow faintly under the pulsating pink and blue lights of towering holographic billboards, while his eyes are hidden behind sleek, futuristic visor goggles radiating a cool violet-blue luminescence that mirrors the city's electric pulse; clad in a high-collared, armored black jacket accented with glowing orange circuitry along its seams, he exudes an aura of stoic intensity and technological prowess, as blurred silhouettes of passersby dissolve into the background, their forms swallowed by the misty haze and shimmering reflections on wet pavement, immersing him in a world where humanity and machine merge beneath the ceaseless drizzle and chromatic glow of a dystopian urban dreamscape.
- **Sample-13**. As the sun dips below the horizon, casting a warm golden glow across the sky that fades into soft blues and purples, Shanghai's iconic Oriental Pearl Tower stands tall and radiant, its spherical sections glowing with pink and purple hues that mirror the twilight, anchoring the city's futuristic skyline against a backdrop of sleek glass skyscrapers and modern high-rises; below, the Huangpu River flows gently, reflecting the fading light and the silhouettes of bridges and riverside trees, while lush green foliage along the embankment frames the scene, adding a touch of nature to the urban grandeur, creating a serene yet dynamic panorama where technological marvels and natural beauty converge in perfect harmony at dusk.
- **Sample-14**. Under a brooding, leaden sky that looms heavy with the promise of storm, a colossal wave rises

in furious majesty—its dark, churning body sculpted by unseen winds into a towering, curling crest that crashes forward in a froth of white foam and spray, its deep indigo and slate-gray depths hinting at the ocean's raw, untamed power; above the tumult, a scattered flock of seabirds soars with outstretched wings, their silhouettes stark against the gloom as they ride the turbulent air currents, embodying both freedom and resilience amid nature's overwhelming force, while the horizon vanishes beneath the swell, leaving only the primal drama of sea and sky locked in eternal, awe-inspiring conflict.

- **Sample-15**. The image showcases a delectable pepperoni pizza presented on a rustic wooden board, set against a dark, textured background that adds a touch of sophistication. The pizza boasts a golden-brown crust with visible char marks from being cooked in a wood-fired oven, indicating a crispy texture. The cheese, melted and slightly browned in spots, blankets the pizza evenly, with some areas showcasing a rich, gooey appearance. The toppings are predominantly pepperoni slices, arranged in a somewhat circular pattern around the edges, while others lie scattered across the surface in various orientations. Each slice of pepperoni is glossy, indicating a fresh, juicy texture, and they are generously placed, making the pizza look hearty and appetizing. Interspersed among the pepperoni slices are small flecks of herbs, likely basil, adding a burst of green color and freshness to the dish. To the right side of the pizza, two fresh basil leaves are artistically placed, their vibrant green hues contrasting beautifully against the warm tones of the pizza and the wooden board. A few more basil leaves can be seen in the foreground at the bottom left corner, scattered more casually than the ones on the pizza itself. There are also a couple of slices of pepperoni lying outside the pizza, further enhancing the visual appeal of the presentation. The overall composition of the image is balanced, with the pizza centrally located, drawing the viewer's attention immediately. The lighting is subtle yet adequate to highlight the textures and colors of the pizza, making it look inviting and mouth-watering. The slight shadows cast by the pizza and basil leaves add depth to the image, creating a three-dimensional feel.
- **Sample-16**. The image depicts a serene night scene at a lively port town. The sky is filled with a bright starry Milky Way galaxy, casting a soft glow over the entire scene. The town features quaint, charming houses with warm yellow lights emanating from their windows, creating a cozy ambiance. At the forefront, there is a group of people gathered around wooden tables, enjoying their time together. They are engaged in conversation and laughter, with cups of coffee or tea in hand. A golden retriever dog sits by one of the tables, adding to the homely atmosphere. To the right, there is a tall streetlight and a small flower arrangement in a pot, further enhancing the quaint charm of the setting. In the background, a harbor is visible with boats anchored, and the town extends with more houses and shops lining the streets, including a bakery sign.
- **Sample-17**. From a high vantage point, the sun rises—or sets—in a blaze of golden-orange light that pierces through a dramatic sky streaked with soft pink, lavender, and deep blue clouds, casting long, ethereal shadows across a vast, snow-blanketed landscape of rolling hills and undulating valleys where a winding road snakes like a ribbon through the serene white expanse; frost-kissed shrubs dot the foreground, their dark branches dusted with snow and catching the warm glow, while the distant horizon fades into a hazy, dreamlike mist, blending earth and sky in a tranquil, almost otherworldly winter tableau that evokes both solitude and sublime beauty beneath the celestial spectacle of dawn or dusk.
- **Sample-18**. The image captures the majestic Forbidden City in Beijing, China, bathed in the warm hues of a setting sun. The scene is dominated by several large, traditional Chinese buildings with elegant, ornate roofs painted in vibrant reds and golds. These buildings feature numerous golden dragons and intricate carvings, typical of imperial architecture. The main structure in the center is an imposing palace with multiple eaves and large golden pillars, its entrance flanked by smaller pavilions. The central building's roof is adorned with intricate patterns and two large, pointed gables, adding to its grandeur. In front of the palace, a wide, open courtyard stretches out, paved with smooth, light-colored stones and bordered by white stone balustrades. These balustrades are decorated with sculpted figures and floral designs, providing a stark contrast to the dark stone of the buildings behind them. The courtyard is devoid of people, emphasizing the serene and historical atmosphere of the site. To the left, more buildings can be seen, each with their own distinct architectural features, though slightly obscured due to the architectural layout. The sky above is a soft gradient from pale blue at the horizon to a warm orange near the sun, which casts a gentle glow over the entire scene. A few wispy clouds are scattered across the sky, adding depth and dimension to the panoramic view. In the foreground, there is a series of white, stone railings and steps leading up to the palace, guiding the viewer's eye towards the impressive structure. The entire area is bathed in the soft, golden light of the sunset, creating a peaceful and timeless quality that highlights the historical significance of this famous landmark.
- **Sample-19**. In this serene, sunset-hued beach scene, a woman stands with her back to the viewer, gazing out at the ocean. She has long brown hair tied loosely behind her head and wears a flowing white sleeveless dress that reaches her ankles. She carries a pair of black flip-flops in her right hand. Her light brown and white dog sits attentively beside her on the sandy shore, their brown and

white fur contrasting with the warm, golden tones of the setting sun. The beach is bathed in the soft, orange glow of the setting sun, casting long shadows and highlighting the texture of the sand. In the distance, the gentle waves roll onto the shore, with the sun's reflection shimmering on the water. To the left, a sailboat sails across the calm sea, its silhouette silhouetted against the warm sky. A wooden lifeguard chair with a red life buoy stands near the center-right of the scene, next to a blanket with a floral pattern draped over its legs. The beach is dotted with footprints, and tall grasses and shrubs frame the scene. A couple of seagulls fly low in the orange sky, adding to the tranquil atmosphere. In the background, a cliff rises, partially obscuring the view, and a few more sailboats are visible on the horizon.

- **Sample-20**. An ancient Greek philosopher is talking on a wireless headset.
- **Sample-21**. A serene elven woman with pointed ears and intricate silver face art gazes thoughtfully, clad in a dark green gown with gold trim. She stands in a mystical, moonlit forest where glowing blue mushrooms illuminate the shadowy trees around her.
- **Sample-22**. The image depicts a small, well-lit home office setup in a cozy room with beige carpeting. The primary focus is a compact wooden desk positioned against a pale wall. The desk has a simple, light-colored finish and is supported by two metal legs, which appear to be adjustable for height. On the desk, there is a black keyboard and a laptop computer on the right side, along with a closed, black-framed flat-screen monitor to the left of the laptop. A white mouse and a pair of sunglasses rest on the keyboard. A single table lamp with a black shade stands next to the keyboard, casting a warm light over the workspace. To the left of the lamp, a small stack of books or papers rests on the desk surface. A black rolling chair with height-adjustable arms is stationed directly in front of the desk. The chair's wheels are visible, indicating its portability. The computer monitor is accompanied by a webcam mounted above it on the wall. Below the desk, the floor is partially covered with a light-colored rug that contrasts with the carpeting. Adjacent to the desk, there is a potted plant with lush green leaves placed on a small round table or stand. The room's background features a bookshelf filled with various books, some of which are visible through open shelves. A white cushioned armchair sits to the left of the desk, suggesting a cozy nook for relaxation or additional seating. On the wall behind the desk, near the corner, a rectangular mirror reflects part of the room, adding depth to the space. An overhead lighting fixture casts a soft yellow glow from above, highlighting the desk area while keeping the rest of the room dimly lit. The overall color palette includes neutral tones—beige, white, and shades of brown—creating a calming and functional workspace atmosphere.
- **Sample-23**. In a vibrant, arid desert landscape bathed in warm, golden hues of sunset, a group of three individuals ventures through a rugged, canyon-like terrain. The woman at the center, dressed in a practical olive-green safari outfit with rolled-up sleeves, khaki pants, and a belt bag slung over her shoulder, walks confidently towards the camera. Her dark hair is tied up in a bun, and she has a focused expression on her face as she gazes at the ground. A small, playful fox stands beside her, attentively looking ahead. The woman's right hand holds a stainless steel water bottle, and her left arm is relaxed by her side. On the right, a man wearing a wide-brimmed straw hat, beige shirt, and cargo pants stands observing the surroundings, while his young son, dressed in an orange t-shirt and black shorts, looks back at them with a curious expression. The man and his son are positioned slightly behind the woman, who appears to be leading the way. In the foreground, a cactus plant with a yellow bloom adds to the desert ambiance. The background features towering red rock formations and sparse vegetation, including a few Joshua trees and desert scrub. A large eagle soars high above, its wings spread wide against the backdrop of a sky painted with swirling clouds in shades of orange, pink, and purple. The sand beneath their feet is dotted with footprints, suggesting they have been walking for some time. The entire scene is imbued with a sense of adventure and exploration, set against the timeless beauty of a desert canyon under a dramatic sunset sky.
- **Sample-24**. Please generate a realistic image of the traditional Chinese Hotan Jade pendant. The pendant is a round jade brand, with a full color of turquoise. The jade is warm and delicate, and the surface is highly polished but not excessively reflective, showing the oily texture of real jade. A traditional Jiangnan garden landscape painting is carved in relief on the jade plaque: the upper half of the picture shows a group of Chinese style buildings arranged in a staggered manner, with roofs featuring upturned eaves and horsehead walls, and rich details. The buildings are interspersed with delicate elements such as small bridges, flowing water, weeping willows, and rockeries. The overall composition is complex but not messy, presenting a freehand feeling of traditional Chinese painting style. The lower part of the screen is relatively blank, with only winding rivers flowing from bottom to right, enhancing the spatial hierarchy. The pendant is hung on a gray green Chinese woven rope, with a simple and natural knot, tightly woven from multiple strands of fine thread, with a tough texture. It is decorated with a small red coral bead directly above it. The background of the picture is a light gray cloth in the style of physical photography as a reference, which is overall realistic and realistic. The style is modern high-quality still life photography, with clear

composition, soft lighting, and focus on the center of the jade plaque, blurring the background details.

- **Sample-25**. A portrait of profound wisdom and quiet contemplation, this elderly man with a long, flowing white beard and deeply lined face is captured in dramatic chiaroscuro lighting against a dark void, his gaze fixed on something unseen beyond the frame.

More "Image-Text-Image" reconstruction results generated by our method. Here, we provide more visual examples of the "image-text-image" reconstruction pipeline using our method, i.e., our encoder processes the input image, generates the output descriptive caption, and then passes it through our decoder to recover it to pixels. See Fig. 4, Fig. 5, and Fig. 6 for details.

Scaling unified-bench for testing. We have expanded the evaluation by increasing the number of source images to 2000. The corresponding results are reported in Tab. 2, validating the effectiveness of our method.

Method	Average	CLIP	LongCLIP	DINO-v2	DINO-v3
Bagel	80.95	87.57	92.88	80.11	63.24
GPT-Image*	83.69	**91.37**	92.81	**86.41**	64.15
Ours	**84.74**	90.22	**94.49**	85.72	**68.54**

Table 2. '*' indicates the latest GPT-Image version used for comparison.

A deeper analysis of the performance drop on OCR/DU would be beneficial. We attribute the performance drop on OCR/DU to *decoder-side text-rendering bottleneck,* rather than to the limitations of our method. In our setup, the decoder is instantiated with **SD3,** whose limited text generation capability yields a **noisy supervision signal** for the understanding model, thereby degrading OCR/DU performance. To verify, we replace SD3 with Qwen-Image (with stronger text rendering). As shown in Tab. 3, the performance drop **completely disappears.**

Decoder	OCR	DU
Enc. Only	75.0	70.0
SD3	68.8	58.4
Qwen-Image	75.9	70.5

Table 3. Decoder choices for better OCR understanding.

MSE as the reward model. We train the unified system under an image-text-image (I2T2I) pipeline and use text as an intermediate representation; **only semantic content is preserved.** The CLIP model naturally captures this semantic similarity. Instead, **MSE operates at the pixel level** and penalizes low-level variations (lighting, texture, style) that do *not affect semantics.* Given this reason, our main paper utilizes the semantic encoder (CLIP) as the reward model, rather than pixel-level supervision.

The required T2I model/capability at the beginning of training. The **T2I model is expected to possess basic long-instruction following capability,** as our reconstructive RL process progressively produces more detailed intermediate captions. If the T2I model or text encoder is constrained by short token limits (e.g., CLIP-style 77-token encoders), it can bottleneck reconstruction quality and thus limit effective RL optimization. Importantly, this assumption is *not restrictive in practice,* as modern T2I foundations (e.g., Qwen/Longcat-Image) already show strong capacity for handling long and structured textual prompts, and thus naturally satisfy this and make our method *broadly applicable.*

Limitations and future works. In our experiment, we observe an unexpected decrease in understanding performance on *visual-text recognition related tasks* (Tab. 7 in the main paper), with accuracy dropping by approximately 10% on Document Understanding (DU) and OCR. This likely stems from the well-known difficulty of current image generation models in faithfully rendering text [36], which may introduce misleading reconstruction rewards during RL and hinder the encoder's ability to learn reliable text semantics. This limitation means that the overall benefits of our framework are currently constrained by the weaknesses of the generation module. Improving text reconstruction fidelity is therefore an important direction for future work. Our framework naturally extends beyond the image, as it treats text as a universal representation; the same reconstructive principle could be applied to other modalities such as audio and video.

Original Image

Generated Caption

(by Encoder)

In this Halloween-themed image, a **black cat** with striking **green eyes** sits gracefully on a wooden surface adorned with a rustic ambiance. The cat is **wearing a classic witch's hat**, characterized by its pointed tip and **soft black fabric** with an orange trim along the brim. This hat rests elegantly atop the cat's head, adding to its mystical appearance.

The **cat's fur is long and dense**, giving it a fluffy texture that contrasts beautifully with the matte finish of the hat. Its white whiskers stand out prominently against its dark fur. The cat's **expression is calm and composed**, looking directly at the camera with a gentle gaze.

Surrounding the cat are **three carved pumpkins**, each featuring a traditional jack-o'-lantern face with triangular eyes, a wide smile, and a jagged mouth. The closest pumpkin to the cat has a prominent flame rising from within, casting a warm, flickering light that adds to the festive atmosphere. **The other two pumpkins are placed slightly further away, one to the left and another to the right**.

The background is blurred but reveals a cozy, dimly lit room with a wooden wall, which enhances the warm, cozy feel of the scene. **A lit candle with a flickering flame** is positioned near the right side of the frame, contributing to the Halloween theme. The overall composition of the image combines elements of Halloween tradition with a whimsical twist using the cat, making it both festive and charming.

Generated Image

(by Decoder)

Figure 4. Example of the "image-text-image" reconstruction result by **our encoder and decoder**.

Original Image

Generated Caption

(by Encoder)

In this vibrant and surreal photograph, a pink flamingo **stands atop a pink skateboard**, defying gravity as if taking flight on a miniature skate park. The flamingo, with its elongated neck, **bright yellow beak**, and delicate legs, appears almost lifelike, rendered in pastel shades of pink and white feathers that shimmer slightly under the sun. Its stance is confident.

The skateboard itself is a vivid pink, featuring **yellow wheels–front** and back–with intricate detailing that adds to its playful appearance. The board rests on a **pink surface**, which could be misunderstood as pavement or ground but is more likely a painted platform due to the uniform color and matte texture. This pink area extends across the lower half of the image, providing a striking contrast to the clear, bright pink shadow cast by the flamingo onto the ground.

In the background, the scene unfolds against a vivid blue sky, suggesting a sunny day. **Two palm trees**, their leaves shimmering in the sunlight, frame the image, **one on the left and another on the right**, both slightly out of focus, emphasizing the **flamingo as the central subject**. The yellow wall behind the flamingo is short and segmented, adding a pop of contrasting color to the otherwise monochromatic backdrop. The wall's yellow hue matches the skateboard, creating a harmonious visual connection between the elements. The entire setting exudes a **whimsical, retro vibe reminiscent** of vintage skateboarding culture, with the flamingo adding an unexpected twist to the classic scene.

Generated Image

(by Decoder)

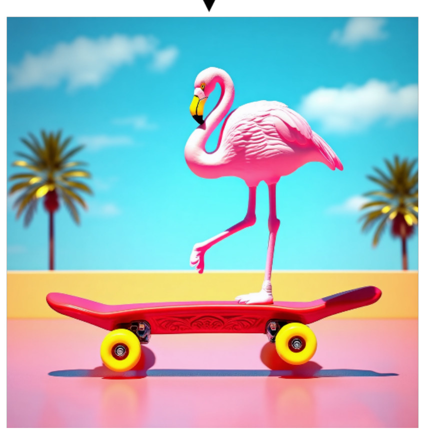

Figure 5. Example of the ”image-text-image” reconstruction result by **our encoder and decoder**.

Original Image

Generated Caption (by Encoder)

In this heartwarming photograph, a **golden retriever puppy** stands with a joyful expression, **gazing slightly to its left**. Its fur is a golden-brown, well-groomed and shiny, with a soft, fluffy texture. The puppy has **small, dark eyes** that sparkle with excitement and a friendly demeanor. Its ears are floppy and slightly darker at the tips, framing its face well.

The puppy wears a **red plaid scarf** that sits snugly around its neck, featuring traditional tartan patterns in shades of **red, black, and white**. The scarf is complemented by a denim jacket with a denim collar, giving it a cozy, winter look. The denim jacket has a casual yet stylish feel, showcasing thick, medium-blue denim fabric with visible stitching and golden metal buttons and snaps. The puppy sports a big, round collar bone on its chest and its **denim jacket features two pockets with matching gold buttons**.

The **puppy's mouth is open**, and its tongue protrudes gently, adding to its adorable charm. Its black nose contrasts sharply against the light fur of its face, and it has a small, dark nose ring just above it.

The puppy stands indoors near a window, bathed in soft, natural light filtering through, creating a warm and cozy atmosphere. Behind the puppy, a window reveals a blurred, sparkling background likely containing snowflakes or blurred lights, suggesting a winter scene. The window itself has white wooden frames, partially obscuring the light. **To the right, a light blue wall provides a soft backdrop**, while the puppy's denim jacket adds a textured element to the image.

Behind the dog, a **blurred green plant leaf peeks into the frame**, contributing to the indoor ambiance. The overall composition is balanced and inviting, emphasizing the puppy's cheerful spirit and the cozy, festive environment it inhabits.

Generated Image (by Decoder)

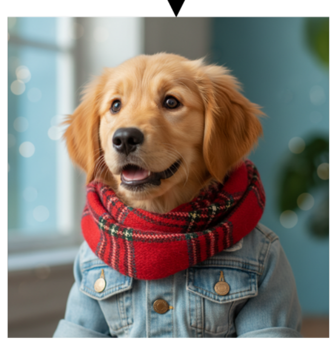

Figure 6. Example of the "image-text-image" reconstruction result by **our encoder and decoder**.

Prompt for the LLM Sampling

System Prompt: You are an expert vision-language model.
User Prompt: Your task is: Given an input image, generate a **textual description** of the image. If there is text in the image, transcribe it inside double quotes.

Now, carefully analyze the input image and output the full description.

Input Image: {{image_path}}

Prompt Used to Perform LLM Judge for Caption Quality

User Prompt:
You will conduct a multi-dimensional analysis of each caption based on the specific criteria listed below. For each criterion, you will assign a score from 1 (very poor) to 10 (excellent). After scoring, you must provide a detailed, structured comparative analysis and declare a final winner.
Evaluation Criteria & Scoring:
Please evaluate each caption against the following four criteria. Provide your scores in a markdown table.
1. **Comprehensiveness, Descriptive Richness, and Accuracy:**
- How deeply does the caption describe the image? Does it go beyond a superficial glance to include important, specific details (e.g., colors, textures, materials, lighting, background elements, expressions)?
- Does it effectively and accurately describe the context (e.g., a black dog not a monkey, or brown eyes not black), environment (background description)?
- Does the caption capture subtle nuances that a casual observer might miss?

2. **Linguistic Fluency and Naturalness:**
- Is the caption grammatically correct and well-written in natural-sounding English?
- Does it flow like a human would describe the scene, or does it sound robotic, disjointed, or like a list of keywords?
- Is the vocabulary choice sophisticated, appropriate, and engaging?

3. **Semantic and Compositional Insight:**
- Does it effectively capture and convey the overall mood, atmosphere, emotion, or narrative implied by the scene?
- Does it demonstrate an understanding of the image's composition (e.g., what is in the foreground vs. background)?

Based on the above rules, provide a comprehensive, head-to-head comparison of the two captions. Structure your analysis with subheadings for each of the four criteria. For each criterion, explicitly quote phrases from both Caption A and Caption B to illustrate your points and justify the difference in their scores. Explain not just *what* is different, but *why* one caption's approach is superior for describing the provided image.

Finally, please declare the winner based on your detailed comparative analysis above. This section must contain only a single letter.

Final Answer: [A or B].

References

[1] Stability AI. Sd3-medium. https://stability.ai/news/stable-diffusion-3-medium, 2024. 5

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 1

[3] Eslam Mohamed Bakr, Pengzhan Sun, Xiaoqian Shen, Faizan Farooq Khan, Li Erran Li, and Mohamed Elhoseiny. Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20041–20053, 2023. 1

[4] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023. 1

[5] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023. 5

[6] Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025. 5

[7] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025. 5

[8] Jaemin Cho, Yushi Hu, Jason M Baldridge, Roopal Garg, Peter Anderson, Ranjay Krishna, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. In *ICLR*, 2024. 1

[9] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017. 1

[10] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. 5

[11] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152, 2023. 1

[12] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *EMNLP (1)*, 2021. 1

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 1

[14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 1

[15] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024. 1, 5

[16] Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 1

[17] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 1

[18] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 5

[19] Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, et al. Holistic evaluation of text-to-image models. *Advances in Neural Information Processing Systems*, 36:69981–70011, 2023. 1

[20] Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Xide Xia, Pengchuan Zhang, Graham Neubig, and Deva Ramanan. Evaluating and improving compositional text-to-visual generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5290–5301, 2024. 1

[21] Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*, 2024. 5

[22] Yi Li, Haonan Wang, Qixiang Zhang, Boyu Xiao, Chenchang Hu, Hualiang Wang, and Xiaomeng Li. Unieval: Unified holistic evaluation for unified multimodal understanding and generation. *arXiv preprint arXiv:2505.10483*, 2025. 1

[23] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*, 2024. 5

[24] Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, et al. Uniworld: High-resolution semantic encoders for unified visual understanding and generation. *arXiv preprint arXiv:2506.03147*, 2025. 5

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

[26] Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025. 1

[27] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024. 1

[28] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. 1

[29] Zichen Miao, Jiang Wang, Ze Wang, Zhengyuan Yang, Lijuan Wang, Qiang Qiu, and Zicheng Liu. Training diffusion models towards diverse image generation with reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10844–10853, 2024. 1

[30] Yuwei Niu, Munan Ning, Mengren Zheng, Bin Lin, Peng Jin, Jiaqi Liao, Kunpeng Ning, Bin Zhu, and Li Yuan. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. *arXiv preprint arXiv:2503.07265*, 2025. 1

[31] OpenAI. Dall·e 3. `https://openai.com/index/dall-e-3`, 2024. 5

[32] Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. Dreambench++: A human-aligned benchmark for personalized image generation. *arXiv preprint arXiv:2406.16855*, 2024. 1

[33] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 5

[34] Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K Du, Zehuan Yuan, and Xinglong Wu. Tokenflow: Unified image tokenizer for multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2545–2555, 2025. 5

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2

[36] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3, 2022. 9

[37] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 1, 2

[38] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 1, 2

[39] Yu Tian, Yue Liu, Shiqi Wang, and Sam Kwong. Quality assessment for text-to-image generation: A survey. *IEEE MultiMedia*, 2025. 1

[40] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 5

[41] Yibin Wang, Zhimin Li, Yuhang Zang, Yujie Zhou, Jiazi Bu, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. Pref-grpo: Pairwise preference reward-based grpo for stable text-to-image reinforcement learning. *arXiv preprint arXiv:2508.20751*, 2025. 1

[42] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025. 5

[43] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13294–13304, 2025. 5

[44] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024. 5

[45] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Wang, Weiyun Ye, Shihao Geng, Yiren Zhao, Jiaming Li, Cunjian Li, Hang Sun, et al. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, 2023. 1

[46] Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, et al. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. *arXiv preprint arXiv:2412.21059*, 2024. 1

[47] Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025. 1

[48] Junyan Ye, Dongzhi Jiang, Zihao Wang, Leqi Zhu, Zhenghao Hu, Zilong Huang, Jun He, Zhiyuan Yan, Jinghua Yu, Hongsheng Li, et al. Echo-4o: Harnessing the power of gpt-4o synthetic images for improved image generation. *arXiv preprint arXiv:2508.09987*, 2025. 1

[49] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. 1